The Analysis & Mining of Globally Distributed Data

Chapter 6. Data Webs

Robert Grossman
Laboratory for Advanced Computing
University of Illinois at Chicago
&
Open Data Partners

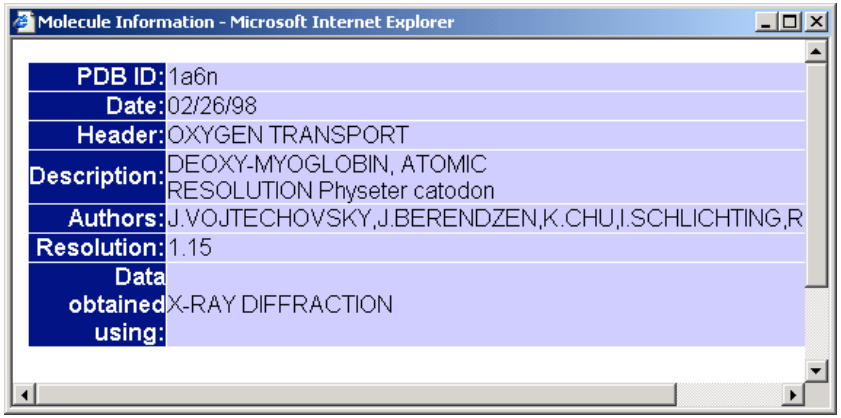
6.1 Data Webs

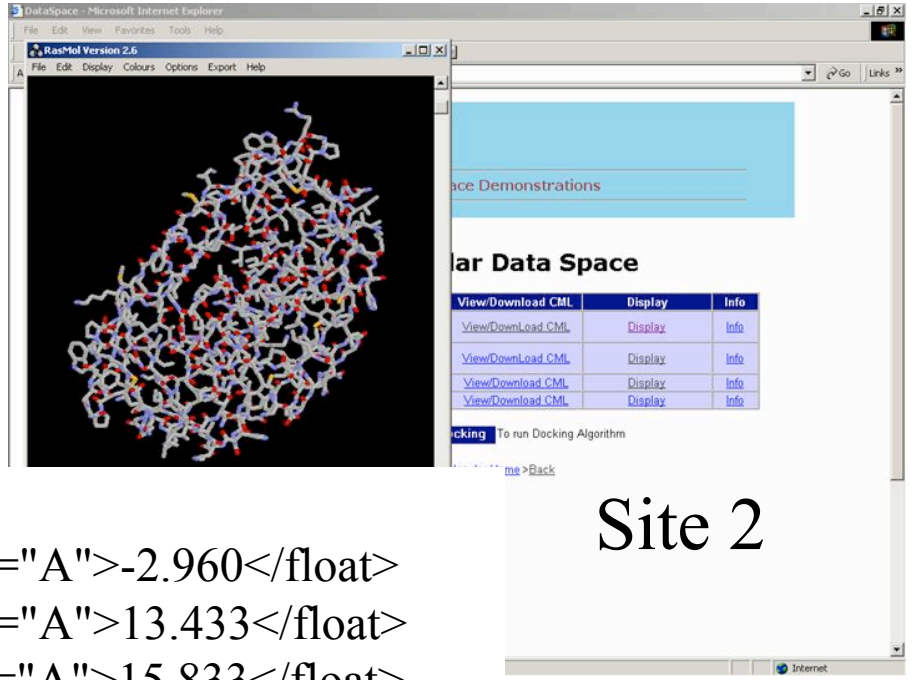
Data meets HTTP.

The Problem

- It is very difficult to use other people's data today.
- To correlate a column of data collected by one person with a column of data collected by another person requires:
  - a graduate student to bring the data to one place and understand the data transformations required
  - several months to get it right

Data Webs – In a Word





Site 1

```
<atom id=4>
   <float builtin="x3" units="A">-2.960</float>
   <float builtin="y3" units="A">13.433</float>
   <float builtin="z3" units="A">15.833</float>
   <string builtin="elementType">O</string>
   <string builtin="residueType">VAL</string>
</atom>
```

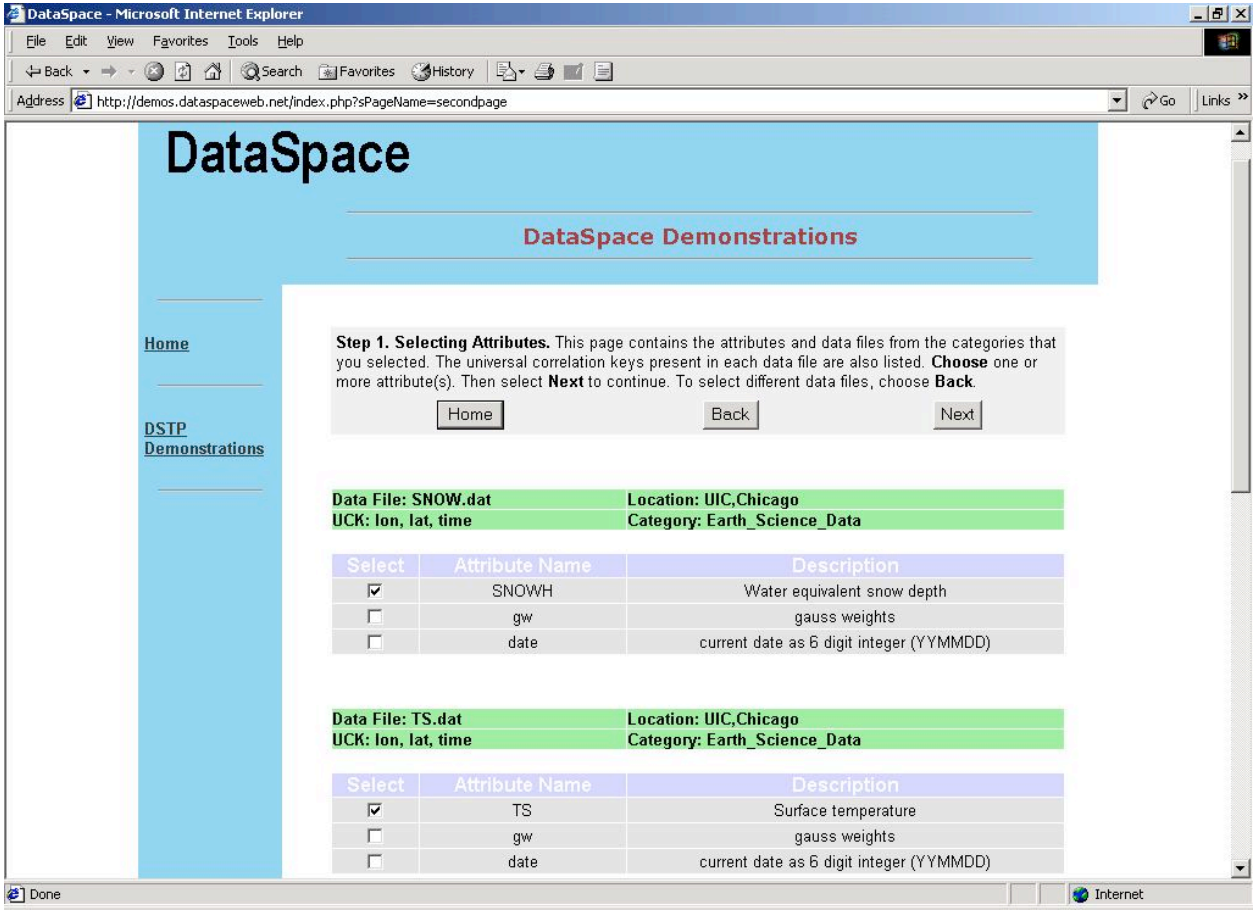
Site 2

- Data just wants to be free.

Key Data Web Protocols & Services

1. Data & metadata selection (DWTP, SQL)
   - using XML metadata, range queries & sampling
2. Data transport (DWTP)
   - DWTP and XML/SOAP
3. Data merging by universal key
   - globally unique UCKs for joining distributed data
4. Data analysis and mining (PMML)
   - using algorithms for clustering, regression, etc.

1. DWTP & SQL – Data Selection



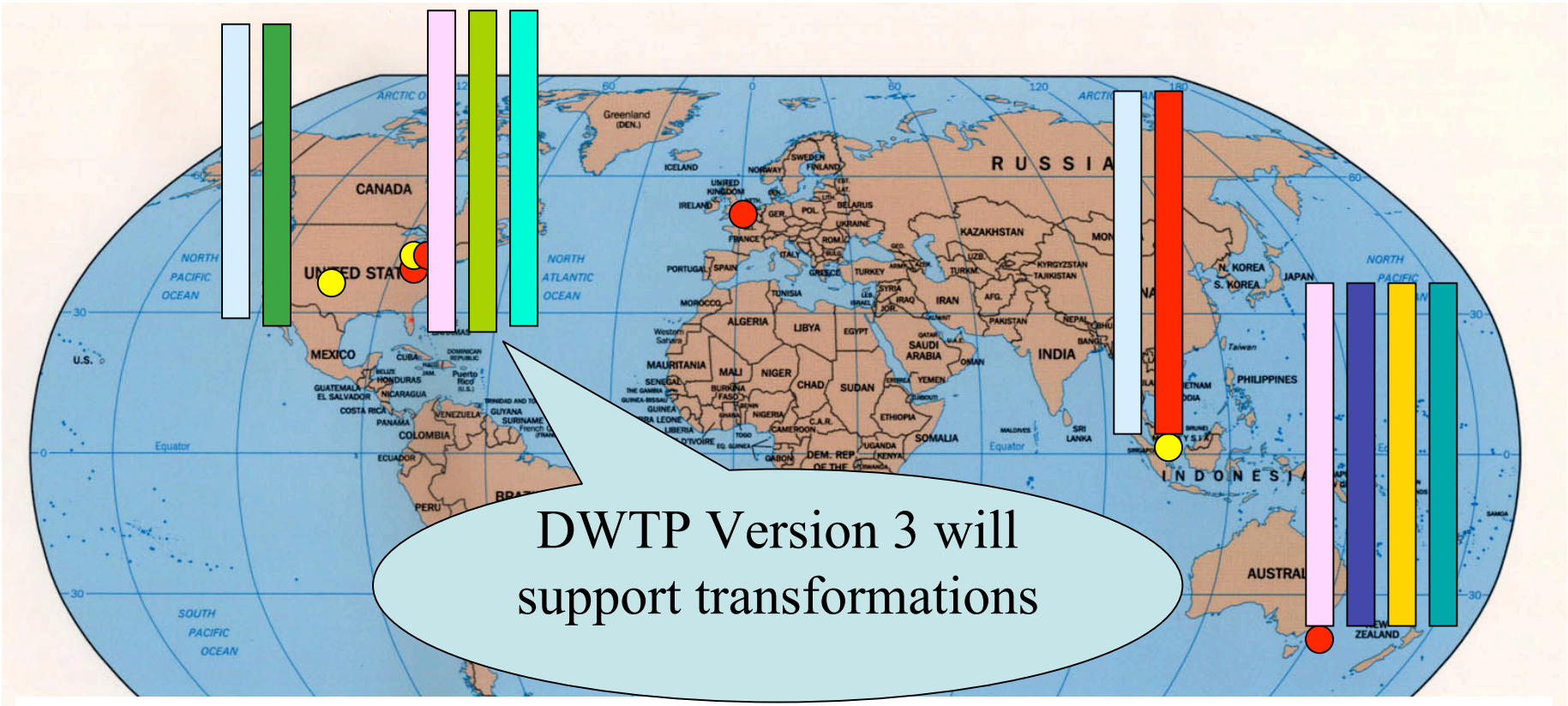
DWTP supports:

- metadata browsing
- selecting attributes
- range queries on records
- SQL
- sampling …

2. DWTP – Data Transport

- DWTP servers support:
  - XML/SOAP services for metadata & small data sets
  - streaming data transport services for data sets and their subsets
- DWTP Servers also support specialized network transport protocols for moving large data:
  - Striped TCP (eg. PSockets, GridFTP)
  - SABUL (reliable UDP)
- DWTP can be 2x – 50x faster than XML/SOAP

3. DWTP – Distributed Joins



View Data as a Collection of Distributed Columns
with Attached Keys (UCKs)

Example: Voting

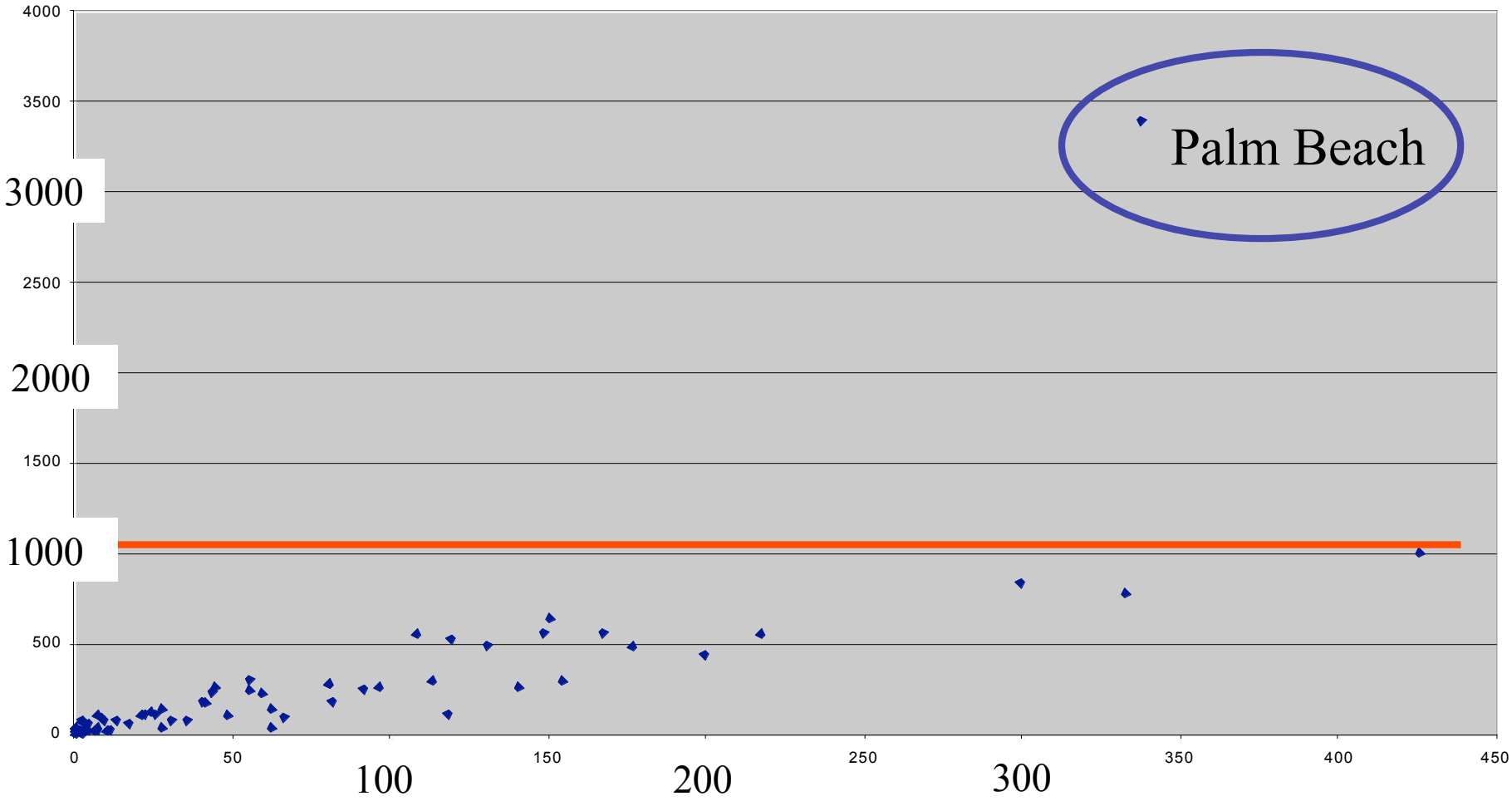
County	Reform
Alachua	91
Baker	4
Bay	55
Bradford	3
Brevard	148
Broward	332

Table 2 – Registered reform voters by county

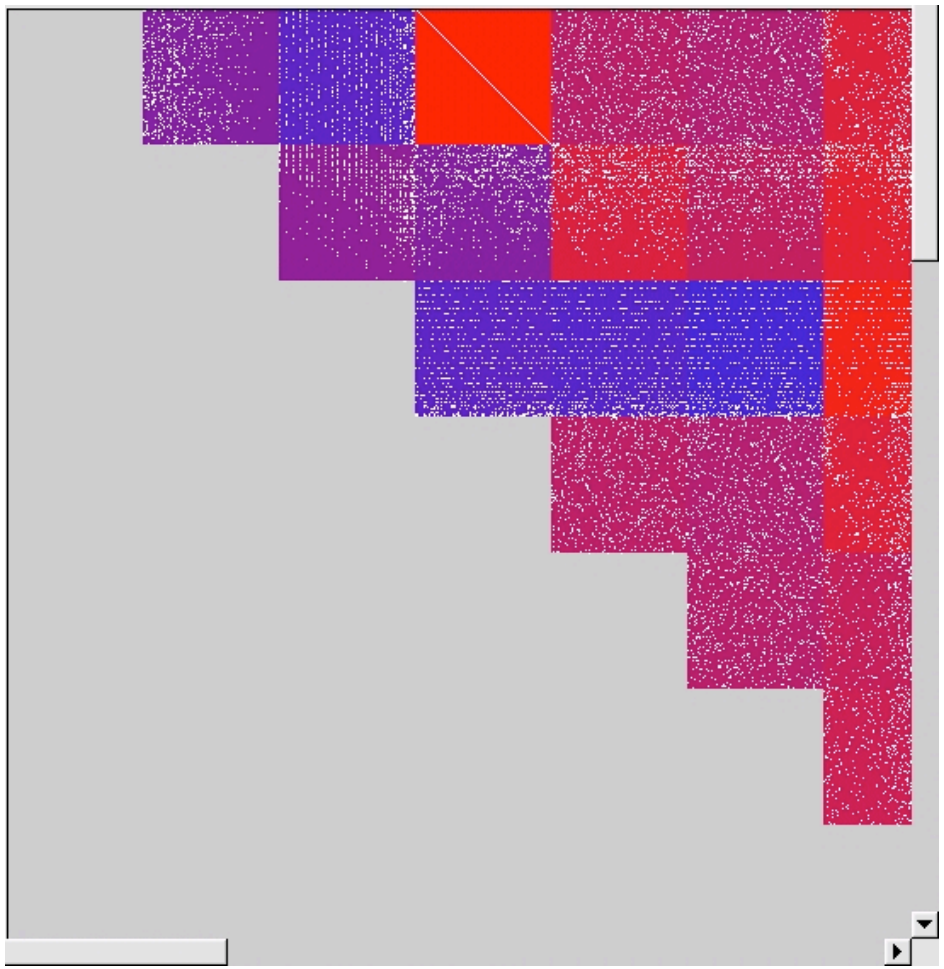
County	BUCHANAN
ALACHUA	263
BAKER	73
BAY	248
BRADFORD	65
BREVARD	570
BROWARD	788

Table 1 – Total votes by county

Correlation: Reform Voters vs Votes for Buchanan



4. Data Analysis & Mining



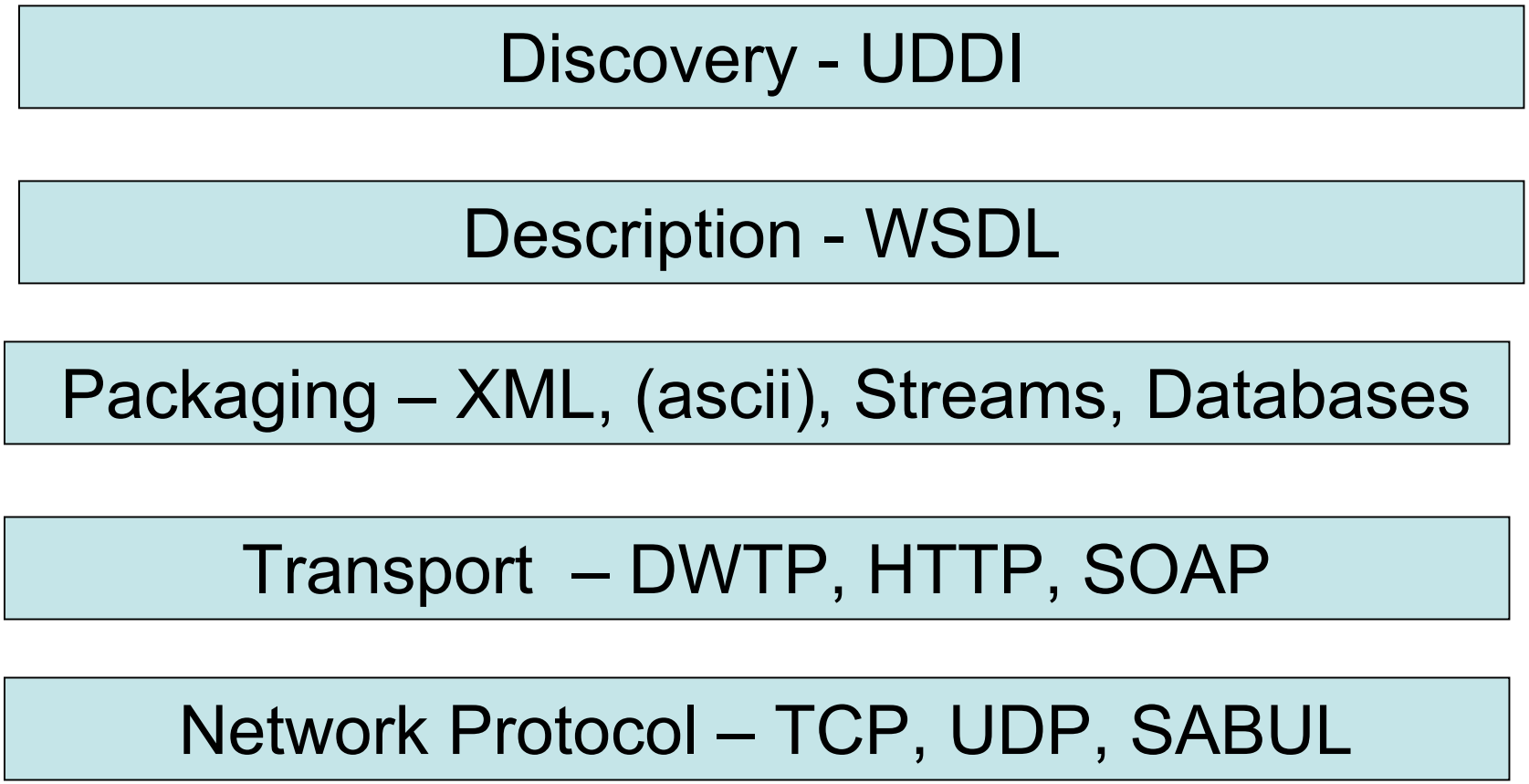
- Data webs often apply data analysis & data mining algorithms to data transported by DWTP
- Usually PMML is produced

DWTP Session

list uck
set uck [uckid]
list datafiles
set datafile [datafileid]
metadata [attributeid]
data [attributeid]
quit

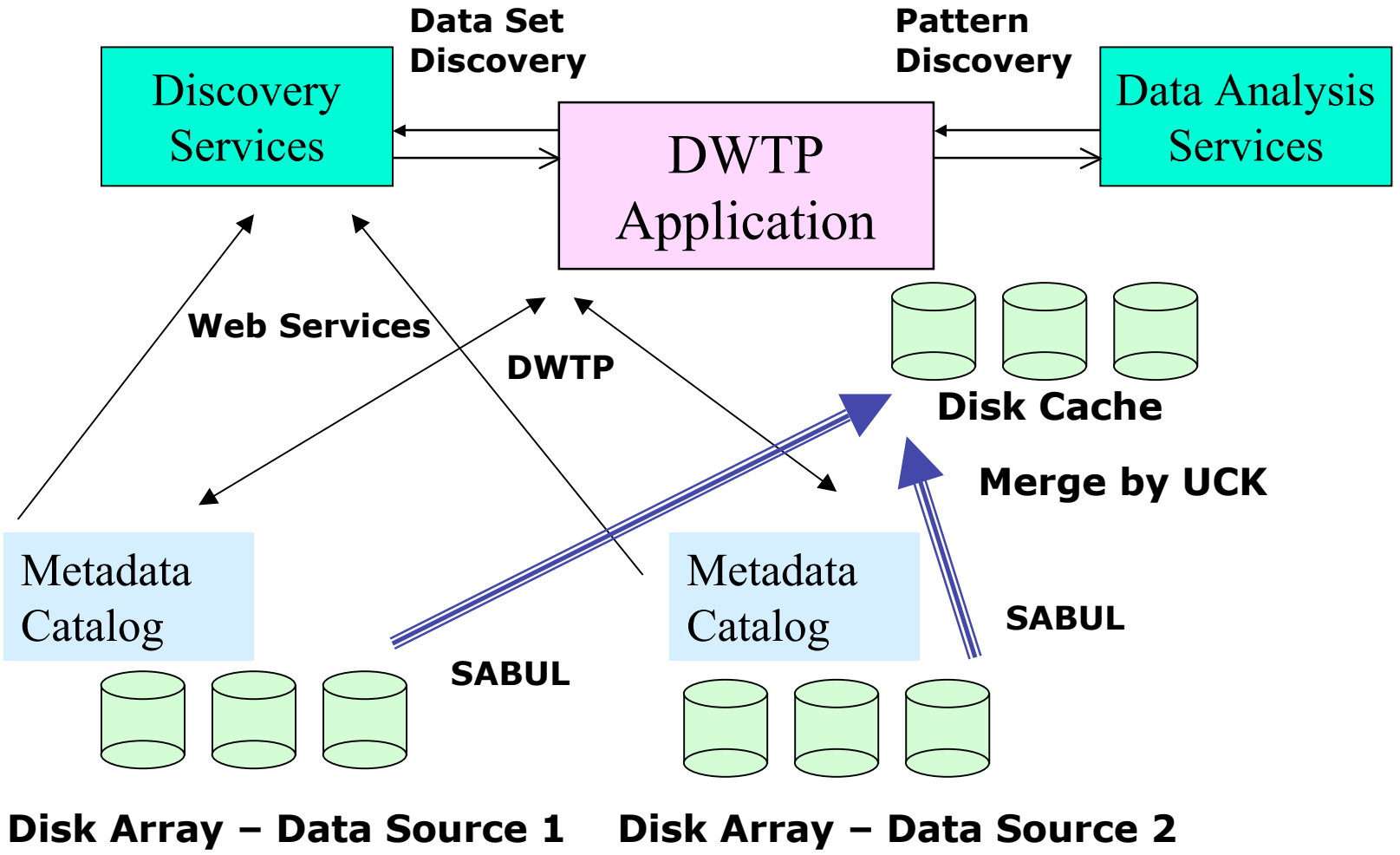
- search for UCK
- for distributed correlation
- which data sets use UCK
- select data set
- what columns do I need?
- retrieve data, merge, & correlate

Stack – Data Web



- Data web model.

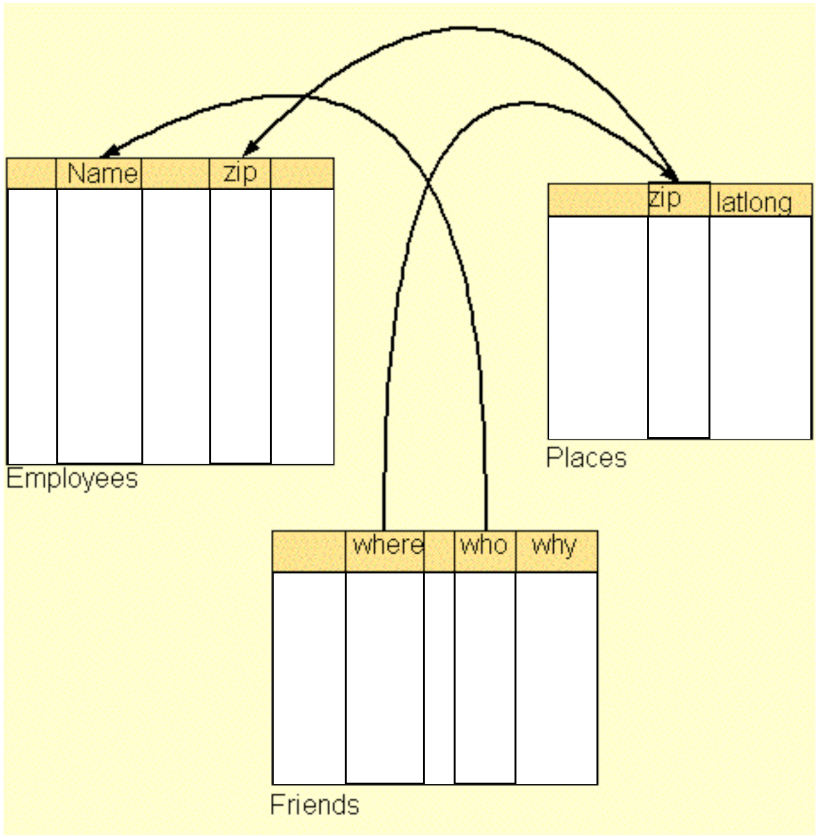
A Model Architecture for Data Webs



6.2 Semantic Webs

From data to knowledge.

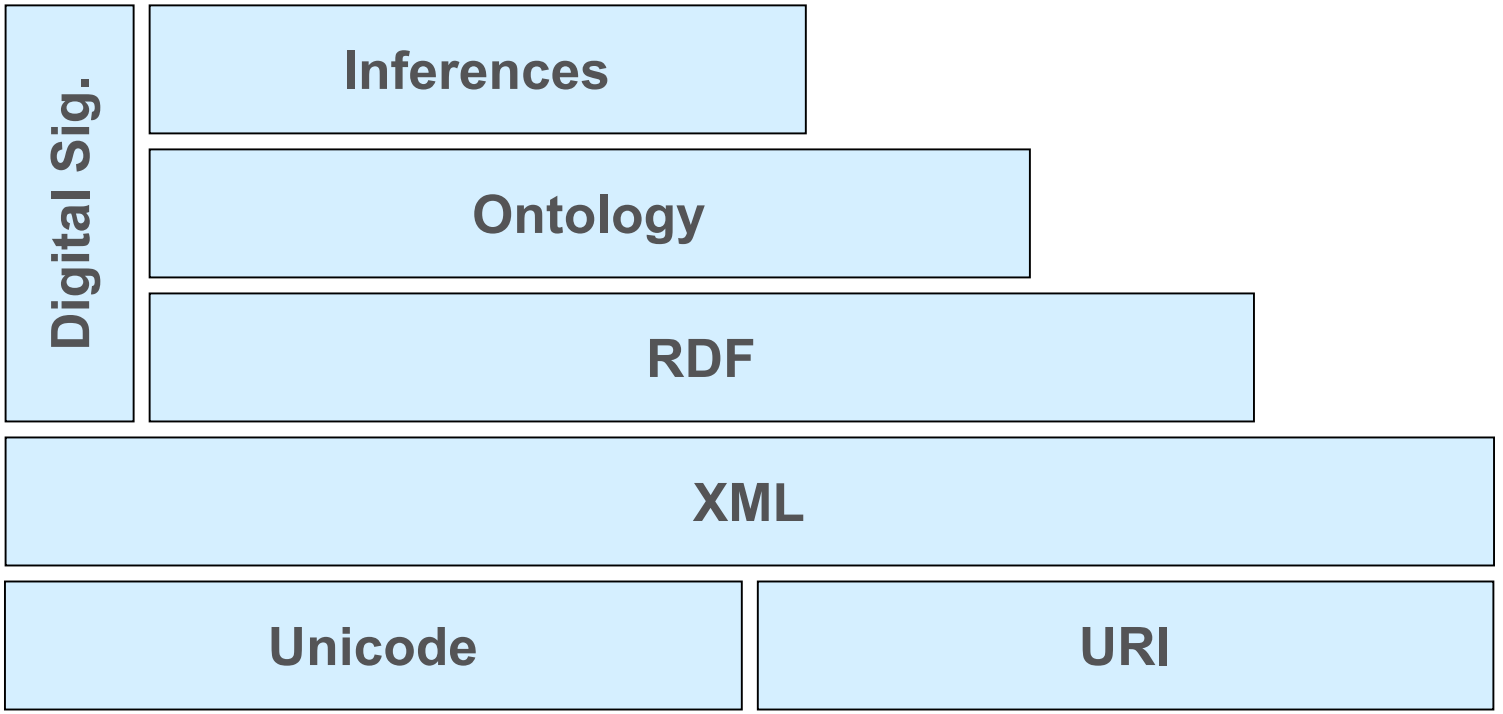
Semantic Webs – In a Word



Source: Tim Berners-Lee, Semantic Web, www.w3.org/2000/Talks/1206-xml2k-tbl/

- It's about knowledge, stupid.

Semantic Web Stack



Source: Tim Berners-Lee, Semantic Web,
www.w3.org/2000/Talks/1206-xml2k-tbl/

Essential Semantic Web Resources & Services

- XML for representing data
- URI for representing resources
- RDF for representing assertions about resources & relationships between them
- Ontologies for mapping between different knowledge domains
- Agents and services to make inferences from RDF & XML data

URI

- URI = Universal Resource Identifier
- URIs are generalizations of URLs which need not have any a network location
- Example: web pages, people, concepts such as “data source”, etc.
- subjects in RDF can be any URI

Resource Description Framework: RDF

- RDF based upon triples:
  - *subject* for the object
  - *predicate* for the attribute or property of the object
  - *object* for the value of the attribute
- Example
  - http://www.dmg.org/pmml-v2.htm has a creation date whose value is October 15, 2002.
  - subject = http://www.dmg.org/pmml-v2.htm
  - predicate = creation date
  - object = October 15, 2002

Ontology

Type	Description	Example
Ontology	Add more complex relationships	Semantic Web
Database schema	Add relationships	DMBS
Programming Languages Class	Add attributes	C++, Java, IDL
Taxonomy	Subclass	Biology

6.3 Summary & References

Data Grids, Data Mining & Data Webs

	Data Grid	Distributed Data Mining	Data Web
Goal	distributed computation	distributed data mining	data explor. & mining
Services	authorization, security, resources	building models, transforming data, etc.	publishing, merging, & correlating columns
Protocol	TCP, GridFTP	TCP	DSTP, …
Platform	dist. clusters	server	dist. cluster

Semantic Web vs. Data Web

	Document Web	Semantic Web	Data Web
Protocol	HTTP	HTTP, SOAP	DSTP, SOAP
Languages	HTML, XML	XML, RDF	XML, PMML …
Action	keyword search	RDF inferences	correlate and mine
Platform	server	server	server, cluster

References – Data Grids

- A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, S. Tuecke, The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets, Journal of Network and Computer Applications, Volume 23, pages 187-200, 2001.
- B. Allcock, J. Bester, J. Bresnahan, A. L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnal, S. Tuecke, Data Management and Transfer in High Performance Computational Grid Environments, Parallel Computing Journal, Volume 28, Number 5, May 2002, pp. 749-771.
- I. Foster, C. Kesselman, J. Nick, S. Tuecke, Grid Services for Distributed System Integration, Computer, Volume 35, Number 6, 2002.
- I. Foster and C. Kesselman, Computational Grids, in The Grid: Blueprint for a New Computing Infrastructure, I. Foster and C. Kesselman, editors, Morgan-Kaufman, 1999.
- A. Chervenak, I. Foster, C. Kesselman, S. Tuecke, Protocols and Services for Distributed Data-Intensive Science, ACAT2000 Proceedings, pages 161-163, 2000.
- R. W. Moore, C. Baru, R. Marciano, A. Rajasekar and M. Wan, Data-Intensive Computing, in The Grid: Blueprint for a New Computing Infrastructure, I. Foster and C. Kesselman, editors, Morgan-Kaufman, 1999.
- S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham, and C. Kesselman, Grid Service Specification, Draft 2, July 17, 2002, Open Grid Service Infrastructure Working Group.

References – Data Webs

- Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, Scientific American, May, 2001.
- Web Architecture: Describing and Exchanging Data, W3C Note 7 June 1999, retrieved from http://www.w3.org/1999/04/WebData.
- Robert Grossman and Marco Mazzucco, DataSpace - A Web Infrastructure for the Exploratory Analysis and Mining of Data, IEEE Computing in Science and Engineering, July/August, 2002, pages 44-51.
- Robert Grossman, Emory Creel, Marco Mazzucco, and Roy Williams A DataSpace Infrastructure for Astronomical Data, in R. L. Grossman, C. Kamath, W. Philip Kegelmeye, V. Kumar, and R. Namburu, Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, 2001, pages 115-123.
- S. Bailey, E. Creel, R. Grossman, S. Gutti, and H. Sivakumar, A High Performance Implementation of the Data Space Transfer Protocol (DSTP), Large-Scale Parallel Data Mining, M. J. Zaki and C.-T. Ho, editors, Springer-Verlag, Berlin, 2000, pages 55-64.
- Reagan W. Moore, Knowledge-based Grids, Proceedings of the Eighteenth IEEE Symposium on Mass Storage Systems, http://storageconference.org/2001/proceedings.html, 2001, pages 29-40.

References – Distributed DM

- S. Stolfo, A. L. Prodromidis, P. K. Chan, JAM: Java Agents for Meta-Learning over Distributed Databases, KDD 1997.
- H. Kargupta, I. Hamzaoglu and B. Stafford, Scalable, Distributed Data Mining Using an Agent Based Architecture, KDD 1997, pages 211-214.
- John Darlington, Yike Guo, Janjao Sutiwaraphun, Hing Wing To, Parallel Induction Algorithms for Data Mining, Lecture Notes in Computer Science, Volume 1280, 1997.
- R. L. Grossman, S. Bailey, A. Ramu, B. Malhi and A. Turinsky, The Preliminary Design of Papyrus: A System for High Performance, Distributed Data Mining over Clusters, in Advances in Distributed and Parallel Knowledge Discovery, H. Kargupta and P. Chan, editors, AAAI Press/The MIT Press, Menlo Park, California, 2000, pages 259-275.
- R. L. Grossman, S. Bailey, A. Ramu, B. Malhi and H. Sivakumar, A. Turinsky, Papyrus: A System for Data Mining over Local and Wide Area Clusters and Super-Clusters, Proceedings of Supercomputing 1999, IEEE.

References – Data Transport

- Robert L Grossman, Harinath Sivakumar and S. Bailey, PSockets: The case for application-level network striping for data intensive applications using high speed wide area networks, Supercomputing, IEEE and ACM, 2000.
- H. Sivakumar, R. L. Grossman, M. Mazzucco, Y. Pan, Q. Zhang, Simple Available Bandwidth Utilization Library (SABUL) for High-Speed Wide Area Networks, submitted for publication, http://www.lac.uic.edu, 2001.
- B. Allcock, J. Bester, J. Bresnahan, A. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel, S. Tuecke, Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing, IEEE Mass Storage Conference, 2001.
- W. Allcock, J. Bester, J. Bresnahan, A. Chervenak, L. Liming, S. Meder, S. Tuecke, GridFTP Protocol Specification, GGF GridFTP Working Group Document, http://www.globus.org, September 2002.
- Robert L. Grossman, Yunhong Gu, Dave Hanley, Xinwei Hong, Dave Lillethun, Jorge Levera, Joe Mambretti, Marco Mazzucco, and Jeremy Weinberger, Photonic Data Services: Integrating Path, Network and Data Services to Support Next Generation Data Mining Applications, Proceedings of the NSF Workshop on Next Generation Data Mining, Kluwer, 2003, to appear.

References – Web Sites

- Data Grids - http://www.globus.org/datagrid
- Semantic Web - http://www.w3.org/2001/sw/ and http://www.semanticweb.org.
- Data Webs - http://www.dataspaceweb.org